HW4 Part 2 - Titanic DataSet Analysis

Eduardo Solano-Salgado

10/04/2020

Summary

The First step in analyzing the Titanic data set is to quickly see the DataFrame and recognize the variables and type of information included. Below is the output of the first 5 rows of data in the DataFrame

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Figure 1: Titanic Head

From the output we can observe that there are 12 different variables, with 2 possible numerical variables (Age and Fare), a primary key (PassengerId), 6 possible categorical variables (Survived, Pclass, Sex, SibSp, Parch, Embarked), and 3 string variables (Name,Ticket, Cabin).

The next step in the dataset analysis involves getting information about the different columns in the DataFrame. Below is the output of the .info() command.

```
Titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Figure 2: Titanic Column Info

From the output, we can see that there are a total of 891 data entries, with null values/missing information in the Age, Cabin, and Embarked columns.

The next step was to obtain a quick statistical summary of all the different variables in the DataFrame. Below is the output of the .describe() command.

```
Titanic.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Figure 3: Statistical Summary

Even though the summary is not very helpful for all of the non-numerical variables, we can still deduce some basic information from the dataset.

Survival Rate and Passenger Class

The first hypothesis I wanted to test was whether there is a correlation between the class of a passenger (1st, 2nd, or 3rd) and survival. In order to get a better idea of the variables, a 2-way table was drawn from the data. The figure below shows the percentage of passengers in each Passenger Class that either died (Survived=0) or lived (Survived =1) in the Titanic.

Survived	Pclass	
0	1	9.0
	2	11.0
	3	42.0
1	1	15.0
	2	10.0
	3	13.0

Figure 4: Survival and Class 2-way Table

The figure below shows the Side-by-side Bar plot of the values seen in the 2-way table, with surivival on the x axis and groups separated by passenger class.

From both the 2-way table and the graph, we can observe that, although for the survival group the percentage of people from each class is almost the same (between 10 and 15%), in the group that did not survive, there

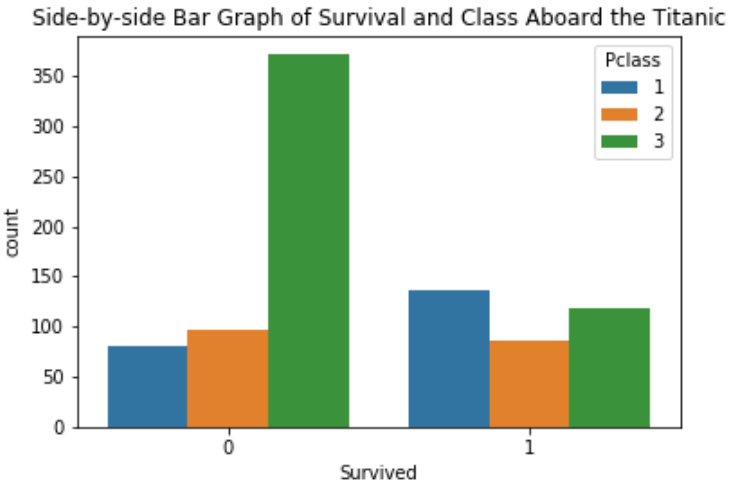

Figure 5: Side-by-Side Graph Survival vs Passenger Class

is a disproportionate amount of passengers from the 3rd class. This implies that there might be a correlation between passenger class and survival rate aboard the Titanic, with a positive bias towards the upper classes. In order to test if there is a significant correlation between the variables, however, we would need to perform a Chi square analysis of the two variables and compare our statistic to a decision point.

Survival Rate and Gender

The second bivariate analysis performed, aimed at testing whether the gender of a passenger had an effect on survival. The first step towards testing this hypothesis involved creating another 2-way table, but with the percentages of men and women who survived or did not survive aboard the Titanic. The figure below shows the 2-way table for gender and survival

Survived	Sex	
0	female	9.0
	male	53.0
1	female	26.0
	male	12.0

Figure 6: Survival vs Gender 2-Way Table

The figure below shows the side-by-side graph of the two variables in the dataset.

From the graph and the 2-way table we can clearly see a correlation between gender and survival aboard the Titanic. We can observe a significantly higher amount of males in the did not survive group and a higher amount of females in the survived group. This implies that gender and survival have a significant correlation among the passangers of the Titanic. Furthermore, this goes with the belief that women and children are generally the first ones to be rescued in a shipwreck. However, just like with the previous hypothesis, the conclusion can only be final once a more complex statistical analysis is performed (such as a Chi square analysis).

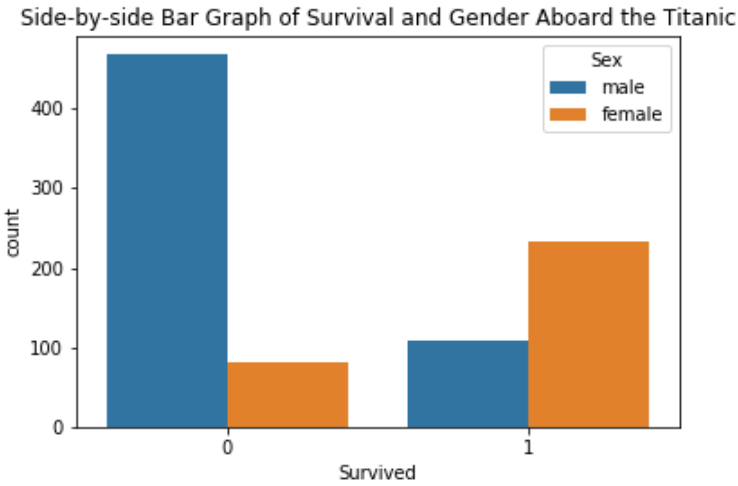

Figure 7: Survival vs Gender Side-by-side Graph

Survival Rate and Age

Finally, I wanted to see if age had any factor in deciding whether a passenger survived or not aboard the Titanic. Since Age is a numerical variable, the first step in performing the analysis of the dataset was to take a look at the distribution of the variable among all passengers. The figure below shows the histogram of the Age variable in the dataset.

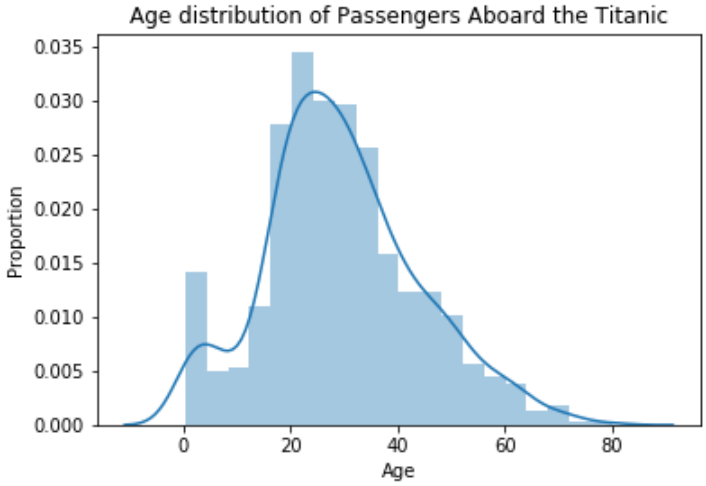

Figure 8: Histogram of Age

From the histogram we can see that age has an approximately normal distribution among the passengers in the Titanic, with a peak centered at around 30. There is also quite a large percentage of young children, as evidenced by a proportion of almost 15% in the first age bin. The next step was to take a look at the age distribution and the statistics of passenger age for each survival category. The figure below shows the .describe() command output of the dataset grouped by their survival categories.

The figure below shows the barplot of the two variables, with survival on the x axis and average Age on the y axis.

	Age							
	count	mean	std	min	25%	50%	75%	max
Survived								
0	424.0	30.626179	14.172110	1.00	21.0	28.0	39.0	74.0
1	290.0	28.343690	14.950952	0.42	19.0	28.0	36.0	80.0

Figure 9: Survival vs Age Statistical Summary

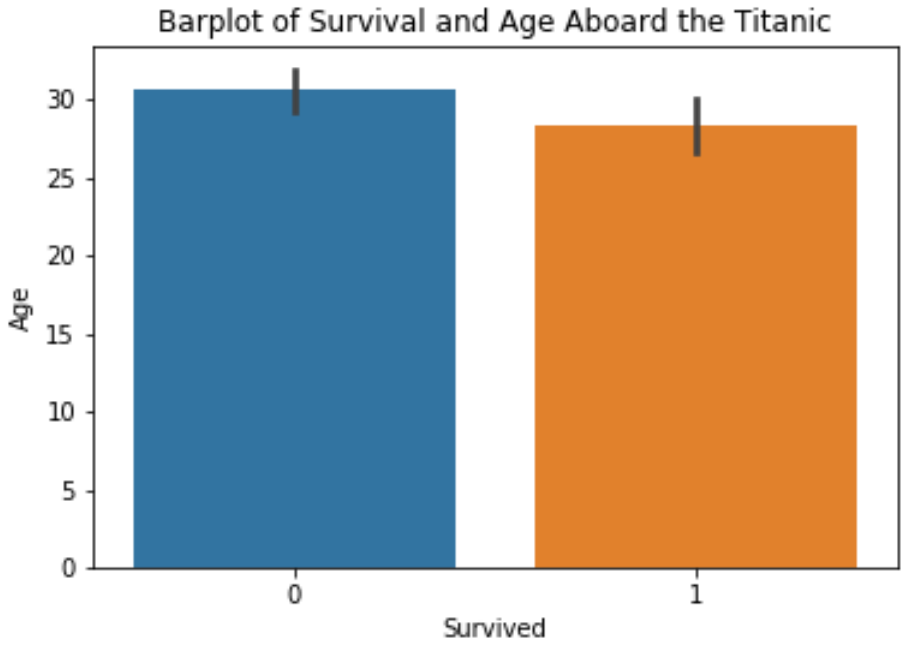

Figure 10: Survival vs Age Barplot

From both the statistical summary table and the barplot, we can see that the average age of the passengers in both survival groups was very simmilar, with the survived group being slightly younger than the did not survive group. Additionally, we can see that for both survival groups, the median age was the same (28) and the standard deviation of the values was very simmilar. To get a better idea of the distribution of age between the 2 survival groups, I compared the histogram of the age variable for each survival group (see figure below).

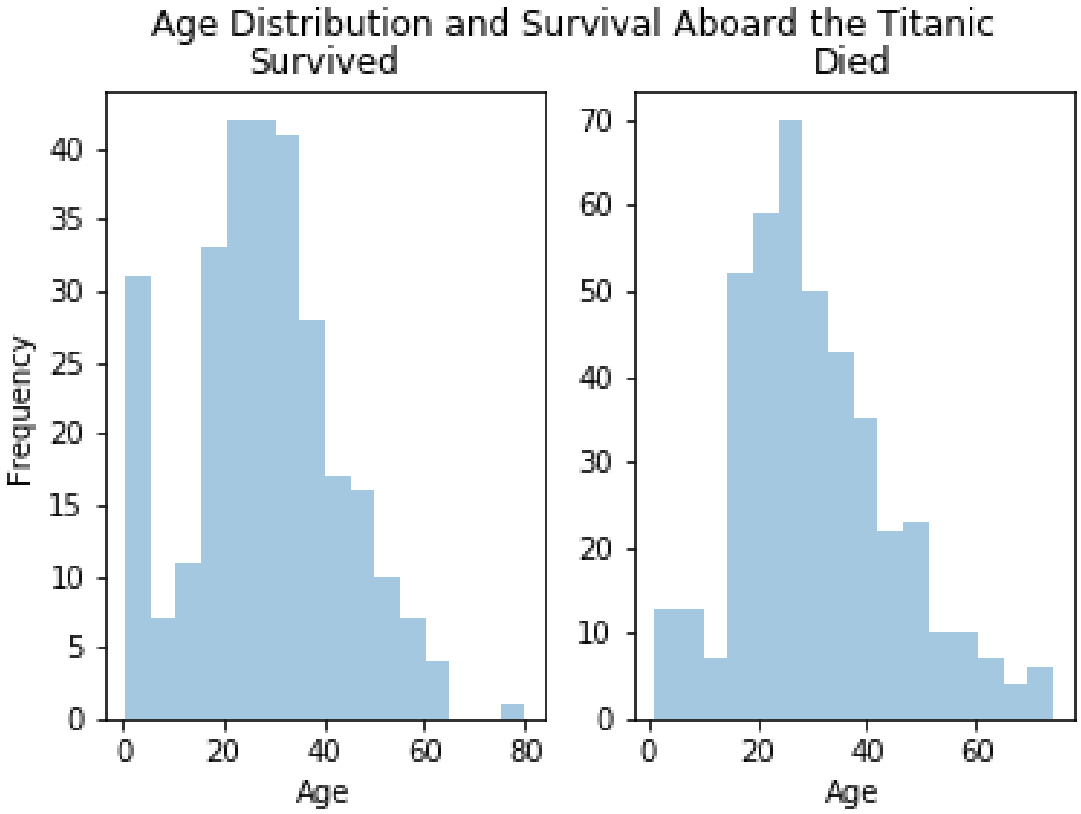

Figure 11: Age Dsitribution Between Survival Groups

From the graphs we can observe that the distribution of age was very simmilar between the survived and did not survive groups. The only major difference, is the larger amount of young children present in the survived group. This also confirms the belief that children are given priority in rescue during a shipwreck.

We can observe that, although young children survived at a higher proportion that the other age groups, age is not significantly related to survival for the majority of age values. Further statistical analyses are needed to understand how age impacted survival rate aboard the Titanic.